Comparative Genome Analysis in the Integrated Microbial Genomes (IMG) System

Victor M. Markowitz[1*] and Nikos C. Kyrpides[2]

[1] Biological Data Management and Technology Center, Computational Research Division
Lawrence Berkeley National Laboratory, 1 Cyclotron Road. Berkeley, USA

[2]Genome Biology Program, Joint Genome Institute
2800 Mitchell Drive, Walnut Creek, USA

*Corresponding Author:
Address: Mail Stop 50A1148, 1 Cyclotron Road, Berkeley, CA 94720
Phone: 510 - 486 7079; Fax: 510 – 486 6363
Email: vmmarkowitz@lbl.gov

Abstract

Comparative genome analysis is critical for the effective exploration of a rapidly growing number of complete and draft sequences for microbial genomes. The Integrated Microbial Genomes (IMG) system (img.jgi.doe.gov) has been developed as a community resource that provides support for comparative analysis of microbial genomes in an integrated context. IMG allows users to navigate the multidimensional microbial genome data space and focus their analysis on a subset of genes, genomes, and functions of interest. IMG provides graphical viewers, summaries and occurrence profile tools for comparing genes, pathways and functions (terms) across specific genomes. Genes can be further examined using gene neighborhoods and compared with sequence alignment tools.

Key Words: comparative genome data analysis, integrated microbial genomes, occurrence profiles

1. Introduction

Microbial genome analysis is a growing area that is expected to lead to advances in healthcare, environmental cleanup, agriculture, industrial processes, and alternative energy. According to the Genomes OnLine Database, about three hundred microbial genomes have been sequenced to date, while over 1000 additional projects are ongoing or in the process of being launched (1). As the genomic community is rapidly moving towards the generation of complete and draft sequences for several hundred microbial genomes, comparative data analysis in the context of integrated genome data sets plays a critical role in understanding the biology of the newly sequenced organisms. Conversely, individual organism-specific genome analysis carried out in isolation cannot support timely analysis of newly released genomes.

Microbial genomes are sequenced by organizations worldwide, follow an annotation process (gene prediction and functional characterization) that is often specific to each sequencing center, and end up in one of the public sequence data repositories, such as GenBank in USA, EMBL in Europe, and DDBJ in Japan.

Genome sequence data include information on gene coordinates, transcription orientation, locus identifiers, gene names and protein functions. Analyzing microbial genomes requires however additional functional annotations, such as motifs, domains, pathways and ontology relationships, which are provided by diverse, usually heterogeneous, data sources, such as Pfam (2), InterPro (3), COG (4), CDD (5), KEGG (6), and Gene Ontology (7). Resources such as EBI Genome Reviews (8) and RefSeq (9) include such additional functional annotations, sometimes after re-annotating the sequences from the public sequence data sources. These resources share common goals, but contain different collections of genomes or data with different degrees of resolution regarding the same genomes. These differences are the result of diverse annotation methods, curation techniques, and functional characterization employed across microbial genome data sources.

Comparative genome data analysis is critical for effective exploration of the rapidly growing number of complete and draft sequences for microbial genomes. For example, the efficiency of the functional characterization of genes in newly sequenced genomes can be substantially improved if this characterization involves methods based on observed biological evolutionary phenomena. Thus, genes with related (coupled) functions are often both present or both absent within specific genomes and tend to be collocated (on chromosomes) in multiple genomes (10). The effectiveness of comparative analysis depends on the availability of powerful analytical tools and the efficiency of the integration, which in turn is determined by the phylogenetic diversity of the organisms, the quality of their annotations, and the level of detail in cellular reconstruction. The efficiency of the integration depends on its breadth (in terms of the number of genomes it involves) and depth (in

terms of different annotations it captures). Integration of available genomic data provides the context for comparative genome analysis, and is becoming the single most important element for understanding the biology of the newly sequenced organisms. Analyzing genomes in the context of other (e.g., phylogenetically related) genomes is substantially more efficient than analyzing each genome in isolation.

The Department of Energy's (DOE) Joint Genome Institute (JGI) is one of the major contributors of microbial genome sequence data, currently conducting about 23% of the reported archaeal and bacterial genome projects worldwide. Individual microbial genomes are sequenced and assembled to draft level at JGI's production facility (PGF), and finished either at PGF, Lawrence Livermore or Los Alamos National Labs. Both draft and finished genomes pass through the automatic Genome Analysis Pipeline (11) at Oak Ridge National Lab (ORNL) which generates gene models and associates automatically predicted genes with functional annotations, such as InterPro protein families, COG categories, and KEGG pathway maps.

Before publication or submission to GenBank, scientific groups interested in a specific genome further review and curate the microbial genome data in collaboration with ORNL's Computational Biology group and JGI's Genome Biology Program. As mentioned above, the efficiency of microbial genome review, curation, and analysis increases substantially when individual microbial genomes are examined in the context of other genomes. Providing such a framework, in order to ensure timely analysis of the genomes sequenced at JGI, is one of the main goals of the Integrated Microbial Genomes (IMG) system (12). IMG aims at providing high levels of data diversity in terms of the number of genomes integrated in the system from public sources, data coherence in terms of the quality of the gene annotations, and data completeness in terms of breadth of the functional annotations.

2. The Integrated Microbial Genomes (IMG) System

The Integrated Microbial Genomes (IMG) system provides support for comparative analysis of microbial genomes in an integrated genome data context. IMG integrates microbial and selected eukaryotic genomic data from multiple data sources. A high level of genome diversity is ensured by collecting data from public sources, such as EBI Genome Reviews, NCBI's RefSeq, and EMBL Nucleotide Sequence Database.

The data model underlying the IMG system provides the structure required for integrating and managing microbial and selected eukaryotic genomic data collected from multiple data sources. The system incorporates in a coherent biological context several data types: (a) primary genomic sequence information, (b) computationally predicted and curated gene models, (c) pre-computed gene relationships (which are sequence similarity based, gene context based, etc.), and (d) functional annotations and pathway information. The user interface is organized in a manner that allows navigation over the microbial genome data space along its three key dimensions representing genomes, genes and functions, respectively.

Genomes (organisms) are identified and organized either based on their taxonomic lineage (domain, phylum, class, order, family, genus, species, strain) or other organism specific properties, such as phenotypes, ecotypes, disease and relevance. For each genome, the primary DNA sequence and its organization in scaffolds and/or contigs, are recorded. Genomic features, such as predicted coding sequences (CDSs) and some functional RNAs, are recorded with start/end coordinates. Predicted genes are grouped based on sequence similarity relationships: ortholog and paralog gene relationships are currently computed based on bidirectional best hit (BBH) single-linkage. COGs provide an additional clustering of orthologous groups of genes in IMG.

Genes are further characterized in terms of molecular function and participation in pathways. Metabolic pathways are modeled in IMG as ordered lists of reactions and consist usually of one to

four reactions. A reaction can include compounds which are reactants (substrates, products) catalyzed by enzymes, and physical entities such as proteins, protein complexes, electrons, etc. Non metabolic pathways are modeled in IMG as lists of functions. Pathways are combined into networks via reactions that share common components. Networks can be further combined into more complex networks. Note that networks are different from KEGG maps which represent complex networks. Pathways are associated with genes via gene products that function as enzymes that serve as catalysts for individual reactions of metabolic pathways. The association of genes with pathways in IMG is based on a controlled vocabulary of terms. IMG terms are defined by domain experts as part of the process of including IMG pathways into the system. The IMG pathways are consistent with the BioPAX (13) level 1 data exchange format in order to facilitate sharing these data across different systems. In addition to the IMG terms and pathways, the GO Ontology is the source for gene functions for the genomes from EBI Genome Reviews, while COGs provide clusters of orthologous groups of genes as further characterization for gene function. Finally, pathways, reactions, and compounds are included from KEGG and LIGAND.

The first version of IMG was released on March 1st, 2005. The current version of IMG (IMG 1.4, as of March 1st, 2006) contains a total of 699 genomes consisting of 395 bacterial, 30 archaeal, 15 eukaryotic genomes and 259 bacterial phages.

3. Comparative Genome Data Analysis in IMG

Data analysis in IMG is set in a multidimensional data space, whereby *genes* form one of the dimensions and are characterized in the context of other dimensions, in particular individual *organisms* (*genomes*), *functions*, and *networks* of *pathways*. Genes are directly associated with genomes (via gene prediction), as well as with functions and pathways (via functional characterization). An organism is associated with a specific function *f* or pathway *p* if its genome has a gene that is associated with *f* or *p*, respectively. Genes can be grouped (clustered) in terms of their sequence similarity or associations with functions and pathways.

Each dimension in the microbial genome data space is characterized by one or several *category* attributes whose values can be used to specify a classification hierarchy. For example, phylogeny serves as a category attribute for organisms and is used to specify their phylogenetic tree classification. Phenotypic attributes, such as origin of the sample used for sequencing (e.g, ocean, groundwater, etc.) can also serve as category attributes for organisms.

Microbial genome data analysis operations allow navigating the multidimensional data space along one or several dimensions and can be set in the context of specific (i.e., subsets of) organisms, functions, and/or pathways. Organism (genome) selections help focus the analysis on a subset of interest, especially in terms of phylogenetic or phenotypic relationships. For example, a set of interest may include all the strains within a specified species. Similarly, function selections focus the analysis on a subset of interest, such as functions involved in lipid metabolism pathways. Finally, gene selections reduce the scope of analysis to genes with certain properties, such as genes sharing a common function or genes that are co-located on the chromosome.

An important type of analysis operation regards examining so called *occurrence profiles* (14, 15) of objects of interest (e.g., functions) selected from one dimension of the multidimensional data space, across objects (e.g., organisms) selected from another dimension of the data space.

Consider two dimensions of the data space representing functions and organisms, respectively. The *occurrence profile* for a function of interest (e.g., enzyme), *f*, shows the pattern of *f* across organisms y_1 to y_n in the form of a vector (L_1, …,L_n) where L_i represents the set of y_i genes that are associated with *f*. Similarly, the profile for a gene, *x*, across organisms y_1 to y_n has the form of a vector (L_1, …,L_n) where L_i represents a set of y_i genes that are associated with *x*, where the association of y_i genes with *x* is based on a specific sequence similarity method.

The number of genes in a set L_i, k_i, is called *gene abundance* and vectors of the form (k_1, …,k_n) are called *abundance profiles*. *Presence profiles* are a special case of abundance profiles, whereby in each vector of the form (k_1, …,k_n), k_i is replaced by either "a" (absent) if k_i is zero or "p" (present) otherwise. Figure 1 shows an example of abundance profiles for genes x_1 to x_4 across organisms y_1 to y_8.

Profiles for objects that are aggregations (compositions) of other objects consist of all the profiles for their component objects. For example, the profile of a metabolic pathway consists of the profiles for the enzymes involved in the pathway, while the profile of a network consists of the profiles of its component pathways.

Analysis based on occurrence profiles usually involves: (i) examining the profiles for objects of a given type across objects of another type; or (ii) finding objects of a given type that either have a predefined presence profile or whose presence profile is similar to the presence profile of a given object of the same type, across objects of another type.

For example, examining the profiles of the genes of a specific organism, y, in the context of other related organisms, y_1,…, y_k allows determining what y may have in "common" with y_1,…, y_k.

Sequences with sufficient degree of similarity are deemed to encode the same gene, and accordingly are considered "common" to or "present" in selected organisms. For the example shown in Figure 1, organism y has gene x_4 in "common" with organisms y_1 to y_8; and genes x_1 and x_2 have the same presence profile across genomes y_1 to y_8. Note that an organism having multiple genes (e.g., three genes of y_4 in Figure 1) corresponding to a specific gene in another organism (e.g., gene x_1 in Figure 1) is the result of the similarity method employed (e.g., homology) in computing profiles. Finding a unique orthologous gene in an organism corresponding to another gene in a different organism is straightforward only for singly copy genes. For other genes, establishing orthologous relationships across organisms is complicated by the fact that most genes undergo either gene duplications or fusion events, with subsequent losses of some of the duplicated copies adding to the complexity of determining such relationships.

Occurrence profile operations can be used for analyzing biological phenomena such as gene *conservation* or *gain*, for a specific organism (e.g., y) in the context of other organisms (e.g., y_1,…, y_k). For the example shown in Figure 1, gene x_4 is conserved across y_1 to y_8, while gene x_3 is gained with respect to y_1 and y_4 to y_8.

Occurrence profiles are critical in the process of understanding the biology of the microbial genome under study. This process is based on observed biological evolutionary phenomena: genes with related (coupled) functions (i) are often both present or both absent within specific genomes that have these functions; (ii) tend to be collocated (on chromosomes) in multiple genomes; (iii) might be fused into a single gene in some genomes; or (iv) are co-transcribed under the same regulator (10).

Consider the example shown in Figure 2, where pathway *p* involves reactions R_1, R_2, R_3, and R_4: genes x_1, x_2, and x_4 of genome G_1 are associated with pathway *p* via enzymes e_1, e_2, and e_4, respectively; genes z_1, z_2, z_3, and z_4 of genome G_2 are associated with pathway *p* via enzymes e_1, e_2,

e_3, and e_4, respectively; if gene x_3 is similar (i.e., determined to be related via significant sequence similarity) to gene z_3, then, following the rules above, x_3 may be associated with *p* via enzyme e_3.

For the example shown in Figure 1, suppose that gene x_1 is functionally characterized while x_2 is not; then the fact that genes x_1 and x_2 have similar occurrence profiles across organisms y_1 to y_8, may help characterize x_2 which may participate in a similar biological process as gene x_1.

Finding objects that have a specific presence profile are used for identifying certain (e.g., *unique*) genes in an organism in the context of other organisms. For example, consider finding genes of a target organism in terms of presence or absence of homologs (or orthologs) in other reference organisms. Reference organisms can be defined based on some biological property, such as phylogenetic relationship, shared phenotype or ecological environment. For example, if the reference organisms are phylogenetically related then finding genes that have a specific profile could be used to identify *preserved*, *gained* or *lost* genes. While the *preserved* genes are shared by all organisms in a phylogenetic lineage and therefore are likely to be inherited from the last common ancestor, gene *gain* and *loss* in the target organism (or group of organisms) can be related to the specific adaptation to the ecological environment of these organisms. A potential application of the occurrence profiles is the identification of genes and other genomic properties that can be used to distinguish between different species or strains of the same species of pathogens using a variety of molecular diagnostics tools.

Occurrence profiles involving functions, pathways, and other genomic data are used in comparative analysis in a way similar to that discussed above for genes. For example, occurrence or abundance profiles of certain COGs (such as signal transduction histidine kinase, serine/threonine proteine kinase and phosphatase) can provide a broad overview of protein families present or absent in the genomes of interest, while occurrence profiles of Pfam domains found in these proteins could provide additional information on the signals sensed by the proteins.

4. Occurrence Profile Analysis in IMG

Comparative genome data analysis in IMG is set in the context of integrated microbial genomes. IMG allows exploring the microbial genome data space along three key dimensions: genomes (organisms), functions, and genes. Comparative analysis for genomes is provided in IMG through a number of tools that allow genomes to be compared in terms of organism-specific summaries (statistics), genes, and functional annotations. We discuss below in more detail the occurrence profile analysis tools provided by IMG.

4.1 Analysis Context

The *context* for occurrence profile analysis is defined by the set of genomes, genes, and functions of interest selected by the user. By default this context involves all the genomes, genes and functions in the system.

Genome (organism) selections provide the option of focusing the analysis on a subset of genomes of interest, such as strains within a specified species. Genomes can be selected using a keyword based *Genome Search* in conjunction with a number of filters, such as such as phenotype, ecotype, disease relevance, or phylum. Organisms can also be selected from an alphabetical or phylogenetically organized list available in the *Organism Browser*. Genome selections can be *saved* in order to set or reset the analysis context.

Genes can be selected using keyword based gene search, sequence similarity search or gene profile based selection. *Gene Search* allows finding genes based on partial or exact matches to a string of characters in specified IMG fields such as gene name or locus tag. Similarity searches are implemented via BLASTp (protein-vs-protein), BLASTx (DNA-vs-protein), BLASTn (DNA-vs-DNA) or tBLASTn (protein-DNA-vs-DNA-protein). Users can define similarity thresholds and select the target database. Gene profile based selection is provided by the *Phylogenetic Profiler*

which is discussed in more detail below. Gene selections can be *saved* in a gene specific *Analysis Cart* called *Gene Cart* (similar to shopping carts of commercial websites) in order to set or reset the analysis context.

Functional roles of genes in IMG are characterized by a variety of annotations, including their COG membership, association with Pfam domains, Gene Ontology (GO) assignments, and association with enzymes in KEGG pathways. Functional annotations can be searched using keywords and filters, with the selected functions leading to a list of associated genes either directly or via a list of organisms. COG categories and KEGG pathways also can be searched and browsed separately. Function selections can be *saved* in a function specific *Analysis Cart* (e.g., COG Cart, Pfam Cart) in order to set or reset the analysis context.

In summary, the *analysis context* is defined by the set of genomes, genes, and functions of interest selected by the user, where the set of genomes is maintained using a genome list, while genes and functions are maintained using Analysis Carts.

4.2 Occurrence Profile Computation Tools

As discussed in the previous section, occurrence profiles are specified in a two dimensional data space, where one dimension represents a set of genes or functions, x_1 to x_n, whose profiles are computed in the context of the other dimension which represents a set of organisms, y_1 to y_m. The occurrence profile for a gene or function of interest, x, consists of a vector of the form (L_1, …,L_n) where L_i represents the set of genes of y_i that are either (a) similar to x (if x is a gene) or (ii) genes of y_i that are associated with x (if x is a function). Occurrence profile results can be displayed as two dimensional matrices or projected on a phylogenetically organized list of organisms.

We present below several examples of employing IMG occurrence profiles in data analysis together with alternative visual presentations of the profile results.

4.2.1 COG Based Functional Occurrence Profiles Example

The following example illustrates how functional occurrence profiles are used in comparative genome analysis. In this example, such a profile is used to examine the presence of a specific pathway (i.e., CO_2 fixation) in a set of selected organisms, namely in the archaeal class of *Methanomicrobia Archaea*. These organisms can first be selected using IMG's phylogenetic based *Genome Browser* as shown in Figure 3 (i) and then saved in order to focus the analysis context as discussed above.

The first step in one of the CO_2 fixation pathways is catalyzed by a CO dehydrogenase/acetyl-CoA synthase enzyme. A keyword search on expression "CO dehydrogenase/acetyl-CoA synthase" with COG as a filter (see Figure 3 (ii)) retrieves a list of 5 COGs corresponding to different subunits of CO dehydrogenase/acetyl-CoA synthase, as shown in Figure 3 (iii). After these COGs are saved with the *COG Cart* (see Figure 3 (iv)), their occurrence profiles across the *Methanomicrobia* organisms are displayed in a tabular format as shown in Figure 3 (v), with each row displaying the profile of a specific COG across the selected organisms. Each cell in the profile result table contains a link to the associated list of genes and displays the count (*abundance*) of genes in this list. Colors are used to represent visually gene abundance, whereby white, bisque and yellow represent gene counts of 0, 1-4, and over 4 respectively.

In this example, the occurrence profile result suggests that, with the exception of one organism, CO dehydrogenase/acetyl-CoA synthase is present in these organisms which means that they rely on this pathway for CO_2 fixation.

4.2.2 KEGG Based Functional Occurrence Profiles Example

The next example illustrates how functional occurrence profiles can be used for comparing phylogenetically related organisms. In the example shown in Figure 4, occurrence profiles of the enzymes participating in nitrogen metabolism are analyzed across the organisms that belong to the

family of *Bradyrhizobiaceae*. These organisms are first selected using IMG's phylogenetic based *Genome Browser* as shown in Figure 4 (i) and saved in order to reduce the analysis context as discussed above.

Starting with the *KEGG Pathway Browser* (see Figure 4 (ii)), enzymes in the Nitogen Metabolism pathway are selected with the *KEGG Pathway Details* as shown in Figure 4 (iii)). A set of enzymes, including nitrogenase, different versions of nitrate reductase and nitrite reductase, is then saved with the *Enzyme Cart* as shown in Figure 4 (iv). The occurrence profiles of these enzymes across the *Bradyrhizobiaceae* family are displayed in a tabular format as shown in Figure 4 (v), with each column displaying the profile of a specific enzyme across selected organisms. Each cell in the profile result table contains a link to the associated list of genes and displays the count (*abundance*) of genes in this list. Note that the occurrence profile tools in IMG provide two alternative display options (functions vs. genomes and genomes vs. functions) as illustrated in this and previous examples.

In this example, the analysis of occurrence profiles shown in Figure 4 (v) suggests that nitrogen metabolism may be different across these organisms.

4.2.3 Gene Occurrence Profiles Example

The following example illustrates how gene occurrence profiles can be used to examine metal binding in *Shewanella*. First, metal binding related functions are found with IMG's *Function Search* using Pfam or InterPro as filters. For example, Pfam 02805 is associated with a list of genes that include *Shewanella* genes that are related to metal binding. These genes are saved using *Gene Cart*, as shown in Figure 5(i). In this example, the presence profiles for genes are displayed in the form of vectors where each position in the vector corresponds to an organism, as shown in Figure 5(ii): the organisms are phylogenetically ordered to facilitate comparison of closely related organisms. Presence of an ortholog of a gene in a given organism is indicated by a domain letter, 'B' for

bacteria, 'A' for archaea, and 'E' for eukarya, while the absence of the gene is indicated by a dot ('.'). One can mouse over the letter or dot to see the organism name along with its phylum. For the example shown in Figure 5, the occurrence profiles for the *Shewanella* genomes are highlighted (see Figure 5 (iii)).

For a single gene, IMG also provides the *Phylogenetic Distribution Viewer* which presents the abundance profile for that gene across the phylogenetically organized list of organisms. The abundance of the selected gene is indicated by the count of homologous genes at each taxonomic level as shown in Figure 5 (iv).

4.3 Occurrence Profile Selection Tools

Occurrence profiles can be used for finding objects (e.g., genes, functions) that share a specific presence profile across a set of organisms. IMG's *Phylogenetic Profiler* is a tool that allows finding genes in a target organism that share the same gene presence profile, where presence or absence of genes is based on (homologous) gene similarity, with cutoffs used to define the similarity relationship.

In the example shown in Figure 6, the *Phylogenetic Profiler* is used to find genes from a *Burkholderia mallei* strain that have no homologs in a *Burkholderia pseudomallei* strain. Similarity cutoffs can be used to fine-tune the selection. The list of genes with the specified profile are then provided as a selectable list as shown in Figure 6.

The *Phylogenetic Profiler* can be used, for example for finding *unique, common, or lost* genes in the (query) organism of interest compared to a target group of organisms. In the example shown in Figure 6, 548 genes are found to be unique in *Burkholderia mallei ATCC 23344 (B. mallei)* with respect to *Burkholderia pseudomallei K96243 (B. pseudomallei).* As we discuss below, such gene

profile based selections provide the context for analyzing phylogenetically related genomes and reviewing their gene models.

4.4 Interpreting Occurrence Profile Results

Occurrence profile results involve organisms, functional roles (e.g., Pfam families, COGs, enzymes), and sets of genes, each of which can be further examined.

For a set of selected organisms comparative summaries are provided using the *Organism Statistics* as illustrated in the left pane of Figure 7, where summaries for the *Burkholderia mallei* and *Burkholderia pseudomallei* strains mentioned above are presented in the context of other related *Burkholderia* strains. These summaries include the total number of genes and enzymes, and the number of genes with various characteristics, such as genes associated with KEGG pathways, COGs, Pfam and InterPro domains. Such summaries can be configured by selecting the properties that are of comparative interest.

Individual organisms can be further examined using the *Organism Details* that includes various statistics of interest, such as the number of genes in the organism that are associated with KEGG, COG, Pfam, InterPro or enzyme information, as shown in the right pane of Figure 7. For each organism one can also examine the associated list of scaffolds and contigs: for each coordinate range, a *Chromosome Viewer* allows displaying genes colored according to COG functional categories.

Individual COG pathways or general categories can be examined using the *COG Browser* which provides a hierarchical listing of the COG general categories (i.e. Amino acid transport and metabolism) and individual pathways (i.e. Arginine biosynthesis). The *COG Pathway or Category Details* lists the COGs of the selected pathway/category and the number of organisms with genes that belong to these COGs. For a given COG, the “organism counts” are linked to a list of organisms

and their associated “gene counts”. KEGG pathways can be explored in a similar manner using the *KEGG Pathway Details*.

Individual genes can be analyzed using *Gene Details*, as illustrated in Figure 8. A *Gene Information* table includes gene identification, locus information, biochemical properties of the product, and associated KEGG pathways. *Gene Details* also includes evidence for the functional prediction: gene neighborhood, COG, InterPro, and Pfam, and pre-computed lists of homologs, orthologs and paralogs. The gene neighborhood displays the target gene with its neighboring genes in a 25kb chromosomal window, as shown in Figure 8, where the target gene is pointed out by an arrow.

The *Gene Ortholog Neighborhoods*, also shown in Figure 8, includes the gene neighborhood of orthologs of the target gene (pointed out by an arrow) across several organisms: each gene's neighborhood appears above and below a single line showing the genes reading in one direction on top and those reading in the opposite direction on the bottom; genes with the same color indicate association with the same COG group. For each gene, locus tag, scaffold coordinates, and COG group number are provided locally (by placing the cursor over the gene), while additional information is available in the *Gene Details* associated with each gene.

A gene can be also examined in the context of its associated pathways, through links to KEGG maps available on the *Gene Information* table. On such a map, the EC numbers are color-coded and linked to the *Gene Details* for the associated genes, as illustrated in Figure 9 which displays the Purine Metabolism KEGG map for the *Burkholderia mallei* gene shown in Figure 8 (pointed out by an arrow in this figure).

4.5 Gene Model Validation

The following example illustrates how occurrence profile results can assist in gene model validation. Consider the *B. mallei* and *B. pseudomallei* genomes mentioned above. The result of the *Phylogenetic Profiler* indicates that, although *B. mallei* is approximately 20% smaller than B. *pseudomallei* (4764 vs 5855 protein coding genes, respectively), it has 548 unique genes (see Figure 6). This high number of unique genes (over 11.5% of the total number of predicted genes) suggests that a large percentage of the coding capabilities of *B. mallei* is distinct compared to *B. pseudomallei*. However, examining these genes using IMG's *Ortholog Neighborhoods,* as illustrated in Figure 10, suggests that most of the differences in gene content between *B. mallei* and *B. pseudomallei* are due to inconsistencies of the gene models. Detailed analysis of these 548 genes subsequently revealed that:

1. genes BMA3300, BMA3308, BMA3320 and BMA3324 appear as unique in *B. mallei,* although each of them has an ortholog in *B. pseudomallei*; these *B. mallei* genes seem to be unique because their ortholog in *B. pseudomallei* was not identified as a valid gene;
2. genes BMA3286 and BMA3303 in *B. mallei* and BPSL0240 in *B. pseudomallei* are functional genes that were erroneously identified as pseudogenes since they supposedly contain authentic frameshifts or stop codons; analysis of their BLAST hits against orthologs in other *Burkholderia* genomes available in IMG shows that they encode full-length proteins with no frameshifts or stop codons and their identification as pseudogenes was based on the alignment to multi-domain homologs – fusion proteins;
3. gene BMA3290 indicates a gene in *B. mallei* which is longer than all its homologs and is likely to have an incorrect start codon; indeed, analysis of this region and its comparison to the regions of synteny in other *Burkholderia* genomes shows that the start codon of BMA3290 is incorrect; moreover, a gene in a different frame was missed due to erroneous prediction of the gene start.

While *Phylogenetic Profiler* shows that *B. mallei* and *B. pseudomallei* have 10 different genes in this region, in fact there is only a two-gene difference: a transposase in *B. mallei*, which is absent from *B. pseudomallei* and an ortholog of BPSL0240, which is a pseudogene in *B. mallei*. Thus, the comparative analysis of the genes in *B. mallei* and *B. pseudomallei* indicates an up to 90% error rate (either false positive genes in one genome or false negatives in the other genome) in the results due to the difference in gene prediction algorithms used to identify CDSs in these two genomes.

5. Summary

Effective microbial genome data analysis across biological data management systems involves providing support for comparative analysis in an integrated data context. We presented the comparative analysis capabilities provided by the Integrated Microbial Genomes (IMG) system, in particular those that are based on occurrence profiles.

The comparative analysis capabilities in IMG are based on techniques that follow observed biological evolutionary phenomena regarding functional coupling of genes (10). Some IMG tools have similarities to analogous tools in microbial genome data analysis systems such as WIT (16), ERGO (17), MBGD (18), SEED (19), Microbes Online (20), and PUMA2 (21). However, IMG has also a number of unique comparative analysis capabilities. Thus, instead of restricting users to a predefined collection of metabolic pathways compiled from the literature and mostly comprising model organisms, IMG provides users with the opportunity to define their own pathways and functional categories by employing Gene, COG, Enzyme and Pfam Analysis Carts regardless of existing annotations. Such user-defined pathways can be further analyzed using a variety of tools, such as COG, Enzyme and Pfam Profiles, and the Phylogenetic Profiler. These tools were specifically developed in order to enable the analysis of genomes that are poorly characterized, are phylogenetically distant from model organisms, and cannot be analyzed efficiently using traditional pathway databases.

The first version of IMG was released in March 2005, followed by quarterly releases consisting of data content updates and analytical tool extensions. A data warehouse framework was used in developing IMG, and was found to provide an effective environment for developing a system that needs to support the integration and management of data from diverse sources, where data are

inherently imprecise and tend to evolve over time. The data warehouse environment has provided an established framework for modelling and reasoning about genomic data.

IMG data content extensions have focussed on data quality in terms of the coherence of annotations, based on sound validation and correction procedures, as well as corroboration of annotations from other public microbial genome data resources. IMG's occurrence profile tools have proved to be effective in the detection and subsequent correction of annotation errors.

We plan to further enhance the occurrence profile tools in IMG. First, we plan to extend the occurrence profile based selection to include additional biological objects, such as gene clusters (e.g., COGs), enzymes, and chromosomal gene clusters. Note that unlike the profile-based selection of genes, no target organism needs to be selected for functional features such as COGs and enzymes that are common to all organisms. In order to support the selection of chromosomal gene clusters, we plan to extend the content of IMG by pre-computing these clusters. Second, we plan to develop improved occurrence profile viewers in order to increase their usability. For example, we are considering presenting occurrence profile results in a hierarchical (tree) phylogenetic context, which would enhance these tools' ability to support examining biological phenomena of interest, such as gene loss and lateral gene transfer. The existing phylogenetic distribution viewer (see Figure 5 (iv)), lays out the taxonomy of each organism in a text-based format which has expressivity limitations. A more intuitive, and therefore more effective, way to represent this type of information in a phylogenetic context could be based on the 16S ribosomal RNA tree.

IMG will continue to be extended through quarterly updates, whereby it aims at continuously increasing the number of genomes integrated in the system from public resources and JGI, following the principle that the value of genome analysis increases with the number of genomes available as a context for comparative analysis. IMG will also continue to address the needs of the scientific community for comprehensive data content and powerful, yet intuitive, comparative analysis tools.

Acknowledgements

We thank Krishna Palaniappan, Ernest Szeto, Frank Korzeniewski, Iain Anderson, Natalia Ivanova, Athanasios Lykidis, Kostas Mavrommatis, Phil Hugenholtz, Anu Padki, Kristen Taylor, Xueling Zhao, Shane Brubaker, Greg Werner, and Inna Dubchak for their contribution to the development and maintenance of IMG. With their comments and suggestions, Krishna Palaniappan and Iain Anderson helped improve the examples in this chapter. Eddy Rubin and James Bristow provided, support, advice and encouragement throughout the IMG project. IMG uses tools and data from a number of publicly available resources- their availability and value is gratefully acknowledged. The work presented in this paper was supported by the Director, Office of Science, Office of Biological and Environmental Research, Life Sciences Division, U.S. Department of Energy under Contract No. DE-AC03-76SF00098.

References

1. Liolios K, Tavernarakis N, Hugenholtz P, Kyrpides, NC. (2006) The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. *Nucleic Acid Research* **34**, D332-D334.

2. Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., et al. (2004) The Pfam Protein Families Database. *Nucleic Acids Research* **32**, D138-D141.

3. Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., et al. (2005) InterPro, Progress and Status in 2005. *Nucleic Acids Research* **33**, D201-D205.

4. Tatusov, R.L., Koonin, E.V., and Lipman, D.J. (1997) A Genomic Perspective on Protein Families, *Science*, **278**, 631-637.

5. Marchler-Bauer, A., Panchenko, A.R., Shoemaker, B.A., Thiessen, P.A., Geer, L.Y., Bryant, S.H. (2002) CDD: A Database of Conserved Domain Alignments with Links to Domain Three-Dimensional Structure. *Nucleic Acids Research* **30** (1), 281-283.

6. Kanehisa, M., Goto, S., Kawashima, S. Okuno, Y., and Hattori, M. (2004) The KEGG Resource for Deciphering the Genome. *Nucleic Acids Research* **32**, D277-D280.

7. Gene Ontology Consortium. (2004) The Gene Ontology Database and Informatics Resource, *Nucleic Acids Research*, **32**, 258-261.

8. Kersey, P., Bower, L., Morris, L., Horne, A., Petryszak, R., et al. Integr8 and Genome Reviews: Integrated Views of Complete Genomes and Proteoms. (2005) *Nucleic Acid Research* **33**, D297-D302.

9. Pruitt, K.D., Tatusova, T., and Maglott, D.R. (2005) NCBI Reference Sequence (RefSeq): A Curated Non-redundant Sequence Database of Genomes, Transcripts, and Proteins, *Nucleic Acid Research* **33**, D501-D504.

10. Bowers, P.M., Pellegrini, M., Thompson, M.J., Fierro, J., Yeates, T.O., and Eisenberg, D. (2004) Prolinks: A Database of Protein Functional Linkages Derived from Coevolution, *Genome Biology* **5**.

11. Hauser, L., Larimer, F., Land, M., Shah, M., and Uberbacher, E. (2004) Analysis and Annotation of Microbial Genome Sequences, *Genetic Engineering*, **26**, Kluwer Academic/Plenum Publishers, 225-238.

12. Markowitz, V.M., Korzeniewski, F., Palaniappan, K., Szeto, E., Werner, G., et al. (2006) The Integrated Microbial Genomes (IMG) System, *Nucleic Acids Research* **34**, D344-D348.

13. BioPAX. (2006) Biological Pathways Exchange. http://www.biopax.org/.

14. Pellegrini et al. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles, *Proc. National Academy of Science* **96** (8), 4285-4288.

15. Osterman, A., and Overbeek, R. (2003) Missing Genes in Metabolic Pathways: A Comparative Genomic Approach, *Chemical Biology* **7**, 238-251.

16. Overbeek, R., et al. (2000) WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction, *Nucleic Acids Research*, **28**, 123-125.

17. Overbeek, R., et al. (2003) The ERGO Genome Analysis and Discovery System. *Nucleic Acid Research* **31**, 164-171.

18. Uchiyama, I. (2003) MBGD: Microbial Genome Database for Comparative Analysis, *Nucleic Acid Research* **31**, 58-62.

19. Overbeek, R. et al. (2005) The Subsystems Approach to Genome Annotation and its Use in the Project to Annotate 1000 Genomes, *Nucleic Acid Research* **33**, 5691-5702.

20. Alm, E.J., Huang, K.H., Price, M.N., Koche, R.P., Keller, K., Dubchak, I.L., and Arkin, A.P. (2005) The Microbes Online Web Site for Comparative Genomics, *Genome Research* **15** (7), 1015-1022.

21. Maltsev, N., Glass, E., Sulakhe, D., et al. (2006) PUMA2 - Grid-Based High-Throughput Analysis of Genomes and Metabolic Pathways, *Nucleic Acids Research* **34**, D369-D372.

	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8
x_1	*2*	*1*	*1*	*3*	*0*	*0*	*1*	*0*
x_2	*1*	*1*	*2*	*2*	*0*	*0*	*1*	*0*
x_3	*0*	*1*	*1*	*0*	*0*	*0*	*0*	*0*
x_4	*1*	*1*	*1*	*1*	*2*	*1*	*2*	*1*

Figure 1. Abundance Profile Example.

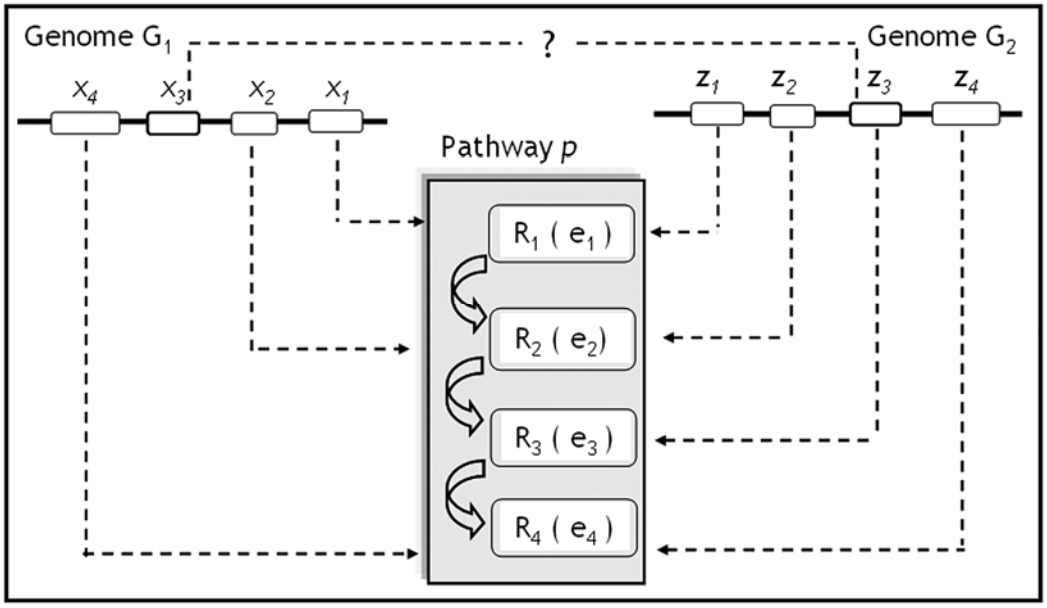


Figure 2. Example of Functional Characterization of Genes.

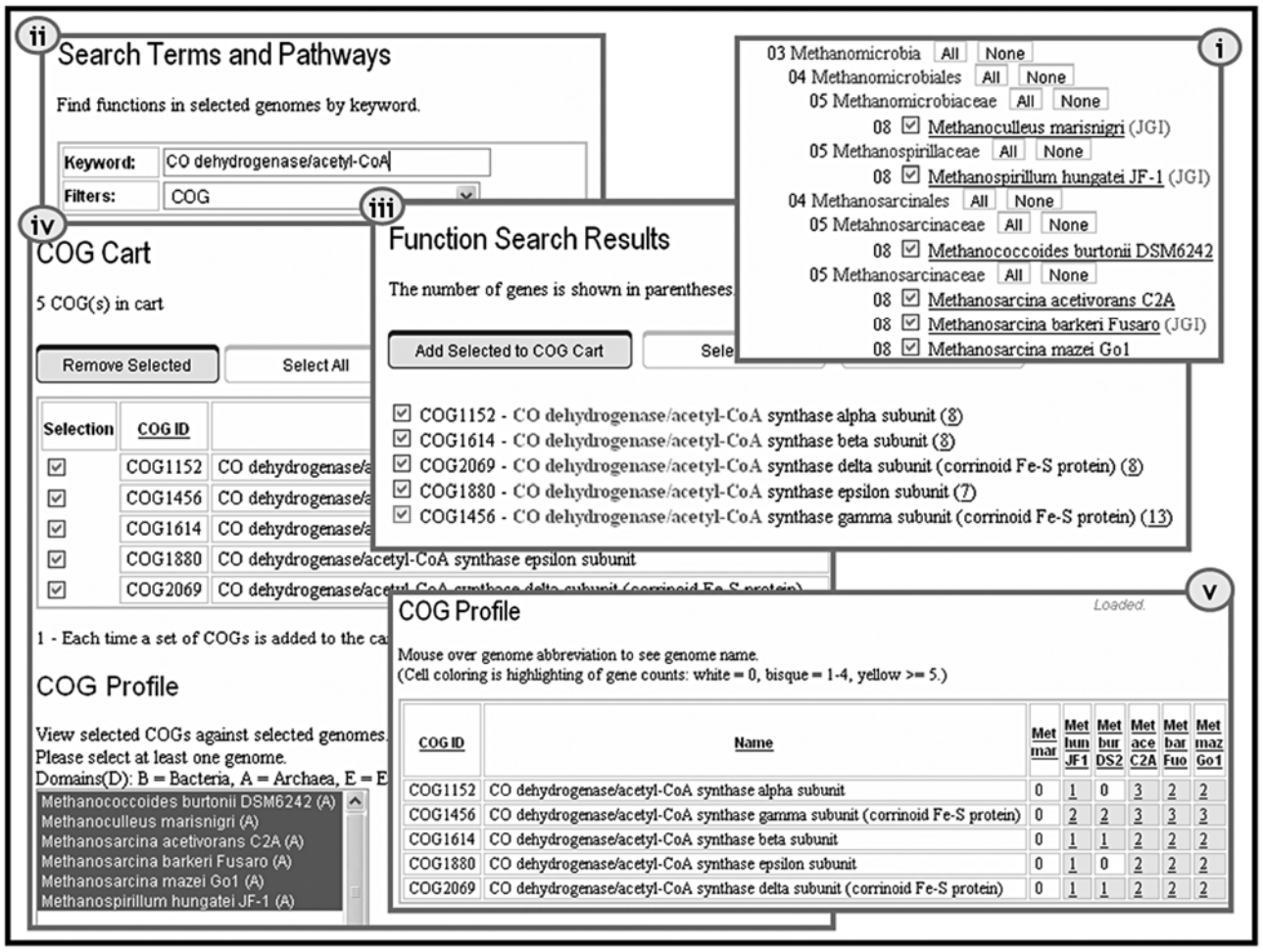


Figure 3. Finding Genes Responsible for Carbon Fixation in *Methanomicrobia Archaea* Organisms.

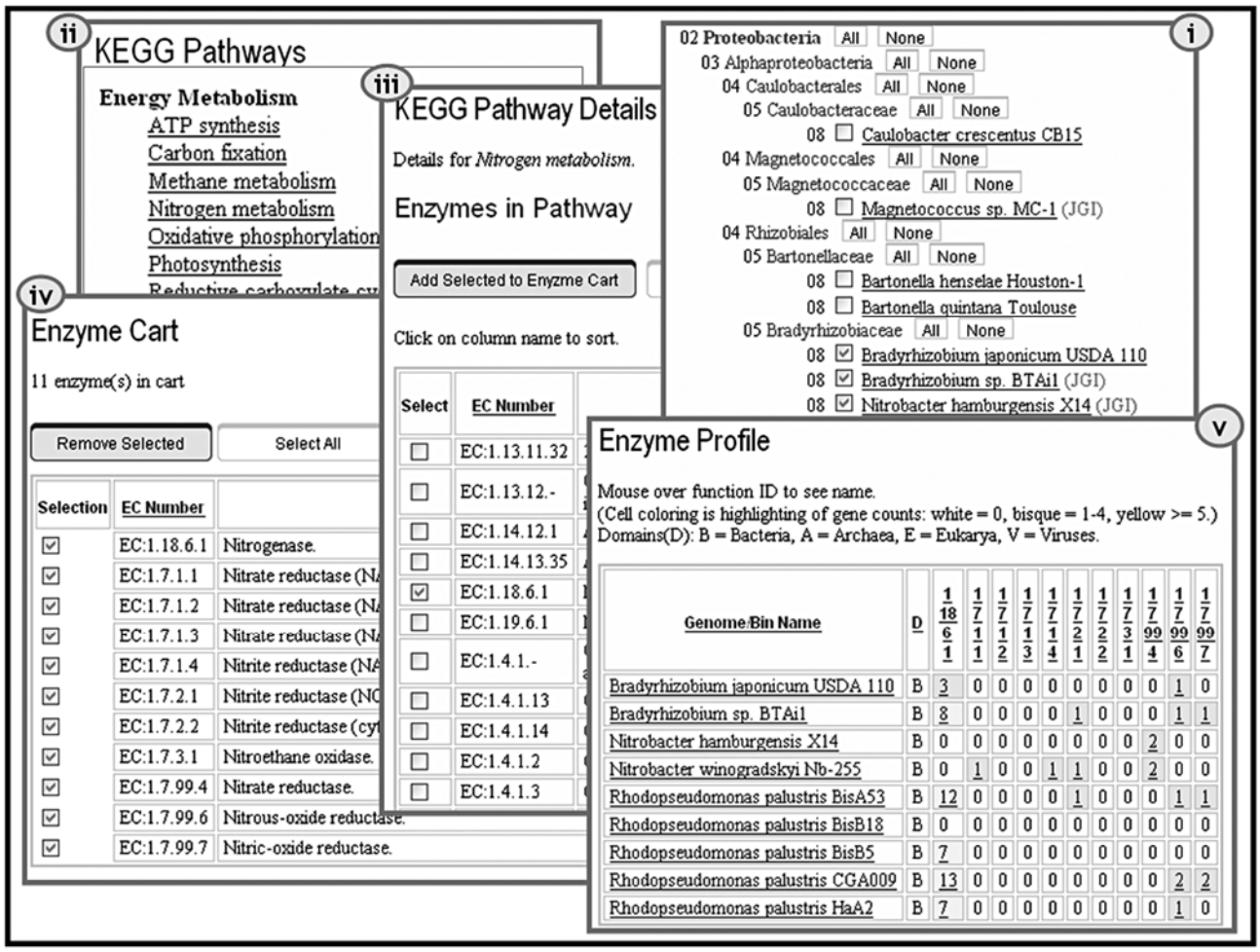


Figure 4. Examining Nitrogen Metabolism in *Bradyrhizobiaceae* Organisms.

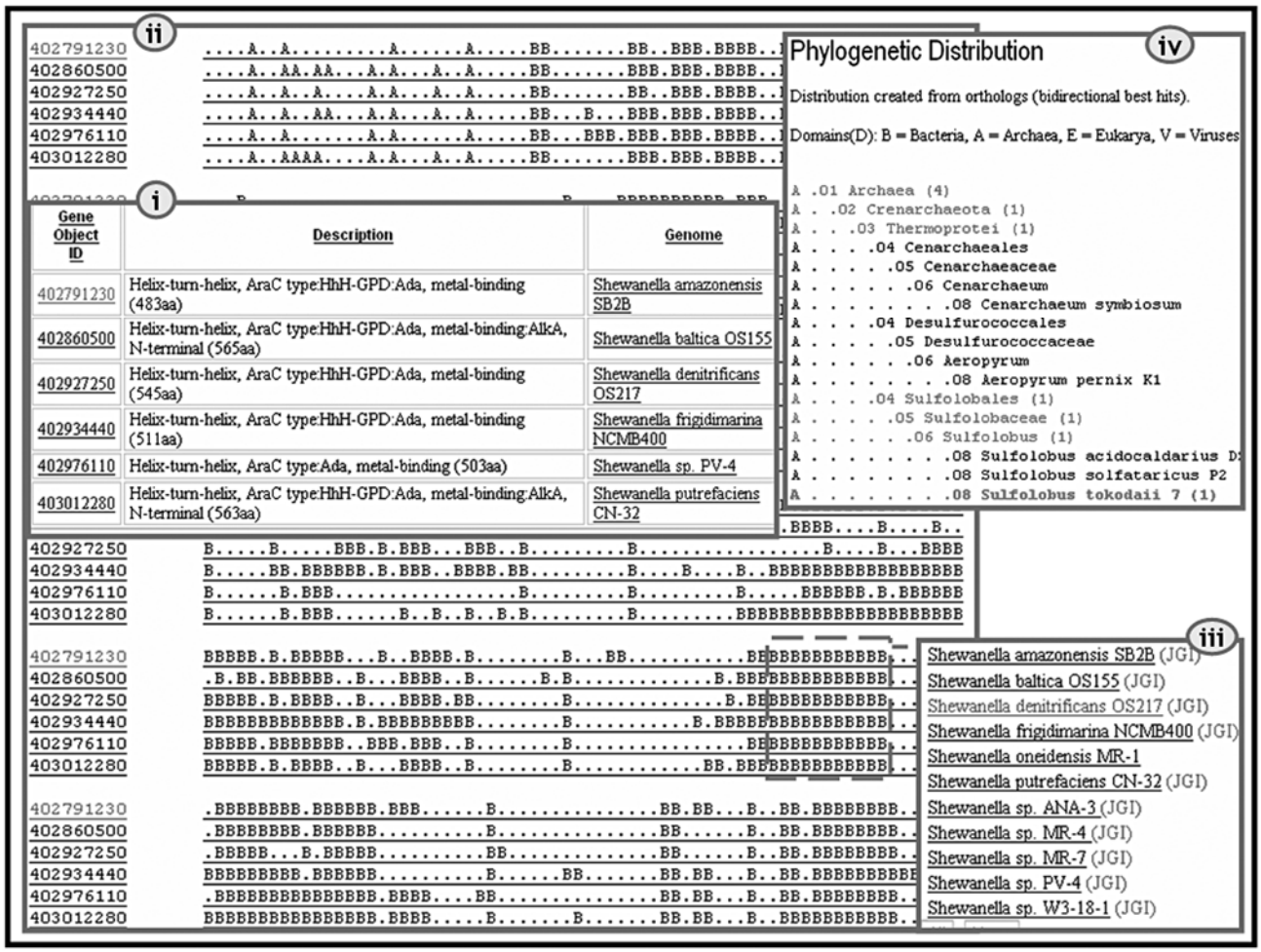


Figure 5. Gene Phylogenetic Occurrence Profile and Distribution Viewer Examples.

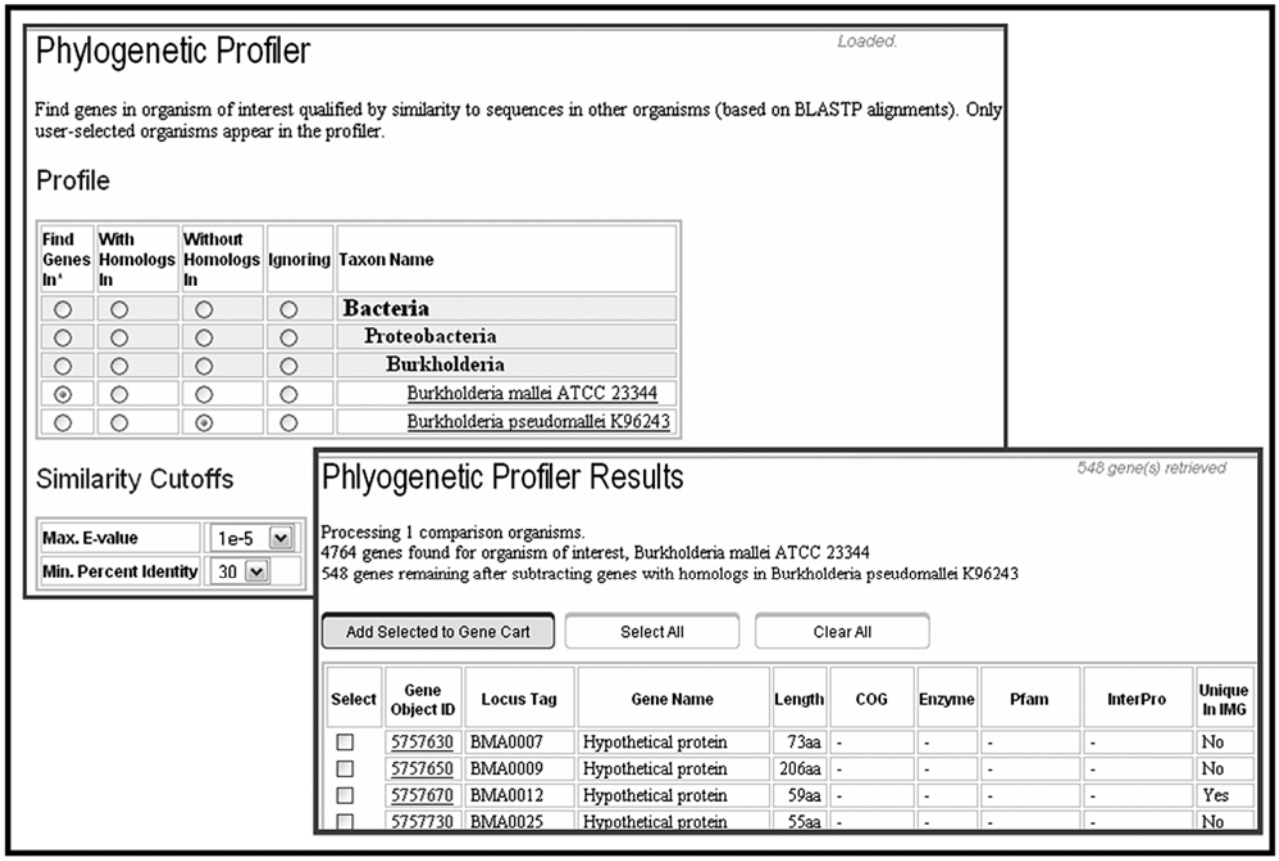
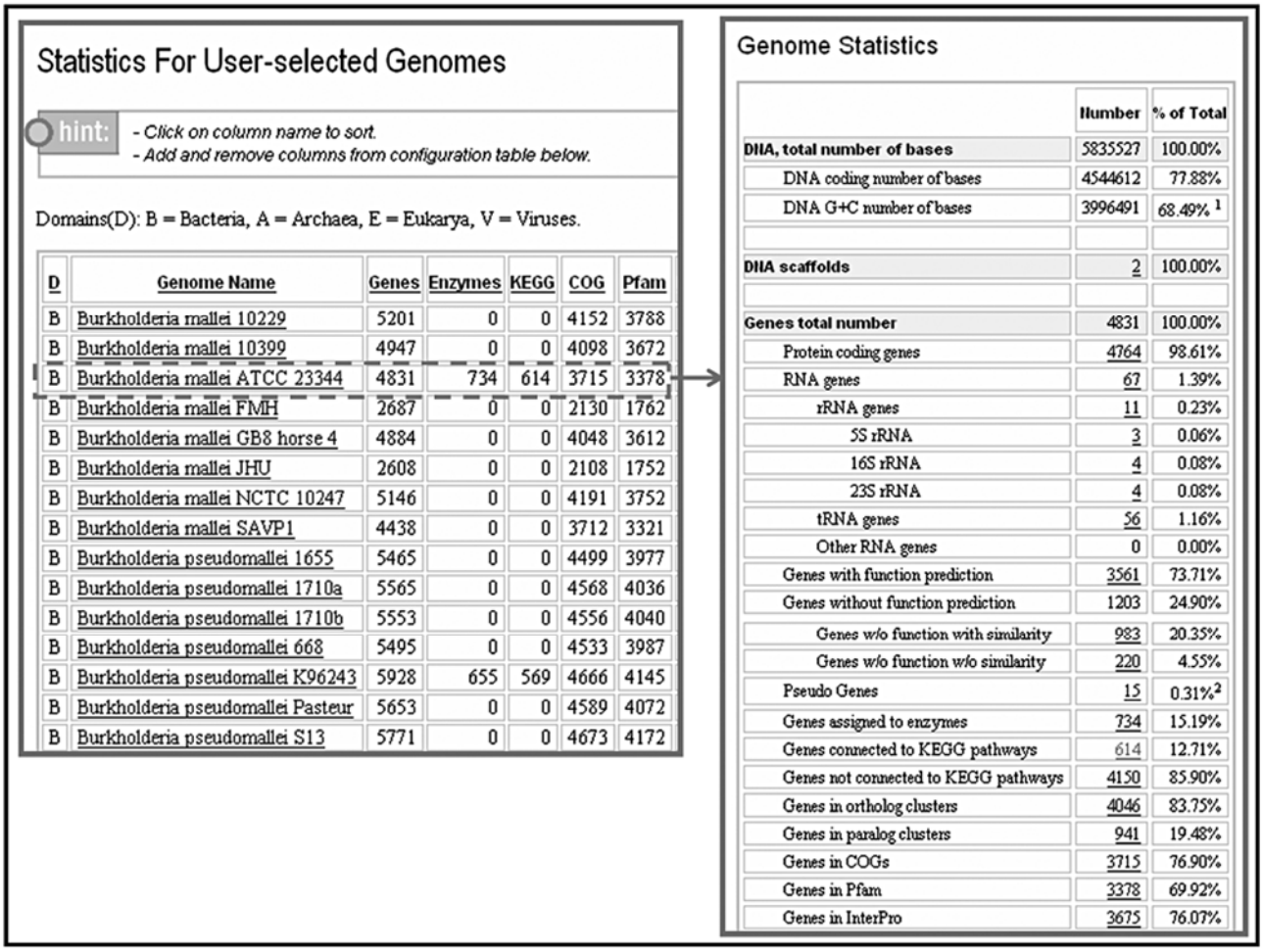


Figure 6. Finding *Burkholderia mallei* Genes Without Homologs in *Burkholderia pseudomallei*.



Statistics For User-selected Genomes

hint: - Click on column name to sort.
- Add and remove columns from configuration table below.

Domains(D): B = Bacteria, A = Archaea, E = Eukarya, V = Viruses.

D	Genome Name	Genes	Enzymes	KEGG	COG	Pfam
B	Burkholderia mallei 10229	5201	0	0	4152	3788
B	Burkholderia mallei 10399	4947	0	0	4098	3672
B	Burkholderia mallei ATCC 23344	4831	734	614	3715	3378
B	Burkholderia mallei FMH	2687	0	0	2130	1762
B	Burkholderia mallei GB8 horse 4	4884	0	0	4048	3612
B	Burkholderia mallei JHU	2608	0	0	2108	1752
B	Burkholderia mallei NCTC 10247	5146	0	0	4191	3752
B	Burkholderia mallei SAVP1	4438	0	0	3712	3321
B	Burkholderia pseudomallei 1655	5465	0	0	4499	3977
B	Burkholderia pseudomallei 1710a	5565	0	0	4568	4036
B	Burkholderia pseudomallei 1710b	5553	0	0	4556	4040
B	Burkholderia pseudomallei 668	5495	0	0	4533	3987
B	Burkholderia pseudomallei K96243	5928	655	569	4666	4145
B	Burkholderia pseudomallei Pasteur	5653	0	0	4589	4072
B	Burkholderia pseudomallei S13	5771	0	0	4673	4172

Genome Statistics

	Number	% of Total
DNA, total number of bases	5835527	100.00%
DNA coding number of bases	4544612	77.88%
DNA G+C number of bases	3996491	68.49% [1]
DNA scaffolds	2	100.00%
Genes total number	4831	100.00%
Protein coding genes	4764	98.61%
RNA genes	67	1.39%
rRNA genes	11	0.23%
5S rRNA	3	0.06%
16S rRNA	4	0.08%
23S rRNA	4	0.08%
tRNA genes	56	1.16%
Other RNA genes	0	0.00%
Genes with function prediction	3561	73.71%
Genes without function prediction	1203	24.90%
Genes w/o function with similarity	983	20.35%
Genes w/o function w/o similarity	220	4.55%
Pseudo Genes	15	0.31% [2]
Genes assigned to enzymes	734	15.19%
Genes connected to KEGG pathways	614	12.71%
Genes not connected to KEGG pathways	4150	85.90%
Genes in ortholog clusters	4046	83.75%
Genes in paralog clusters	941	19.48%
Genes in COGs	3715	76.90%
Genes in Pfam	3378	69.92%
Genes in InterPro	3675	76.07%

Figure 7. Examining Organism Statistics for *Burkholderia mallei and pseudomallei* strains.

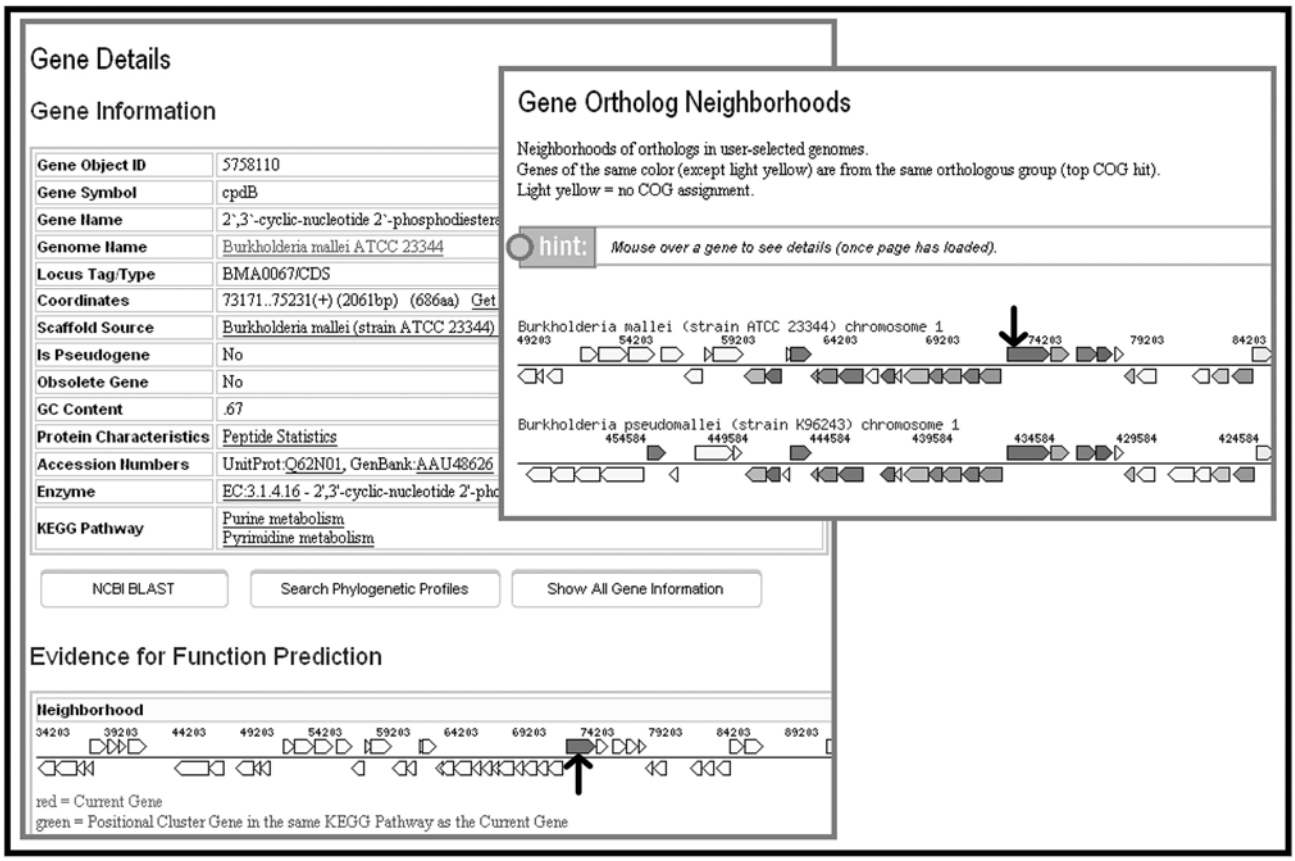


Figure 8. Gene Details and Gene Ortholog Neighborhoods for a *Burkholderia mallei* Gene.

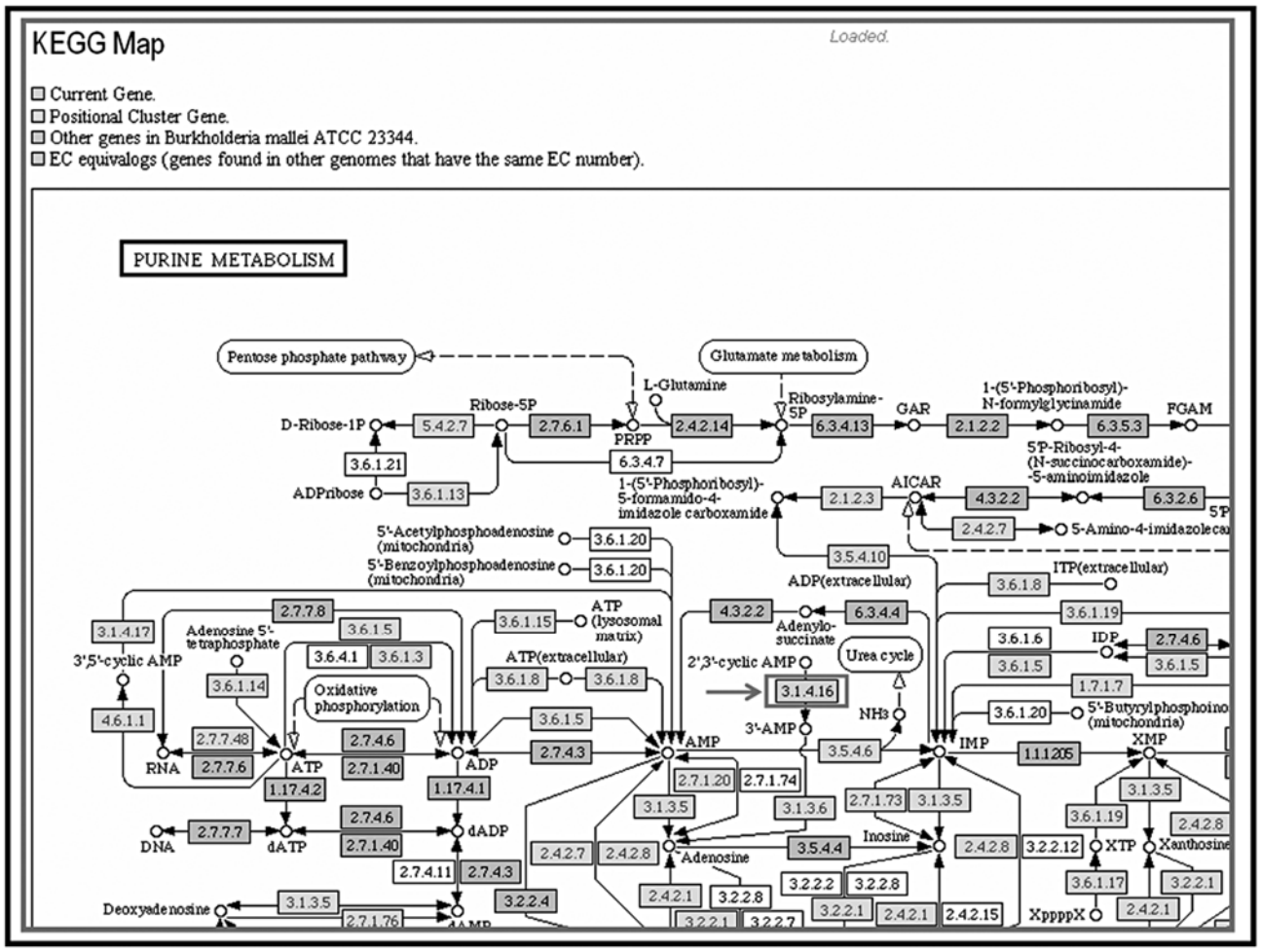


Figure 9. Examining a the Purine Metabolism Map for a *Burkholderia mallei* Gene.

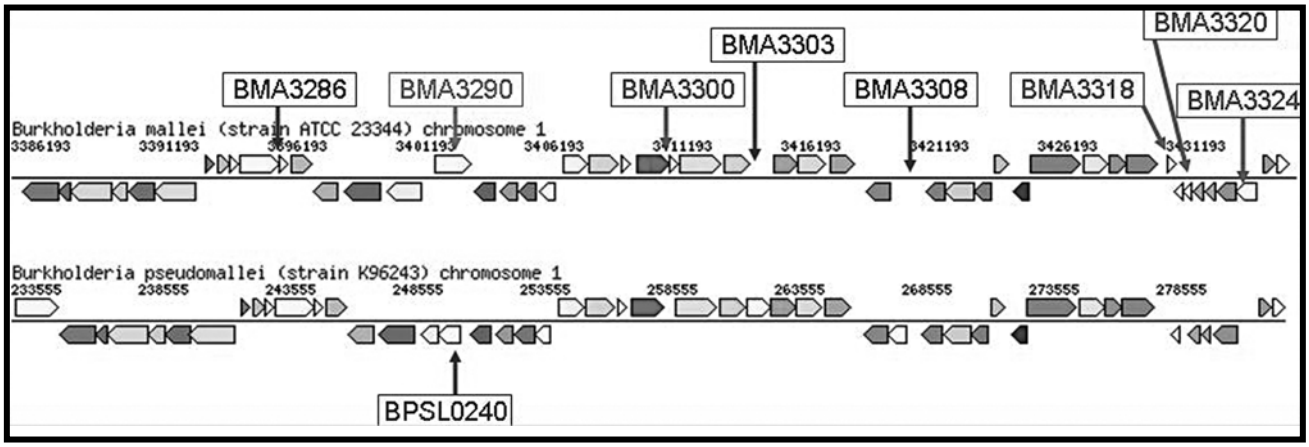


Figure 10**.** Gene Ortholog Neighborhoods for a Region of *Burkholderia mallei* and *Burkholderia pseudomalei.*